图书情报工作
第 65 卷 第 8 期　2021 年 4 月

敦煌壁画叙词表关联数据实体语义相似度计算方法与实验*

■ 高劲松[1]　付家炜[1]　李珂[2]
[1] 华中师范大学信息管理学院　武汉 430079　[2] 青岛海信日立空调营销股份有限公司　青岛 266510

摘　要：[目的/意义] 随着文化遗产数字化和人文计算研究范式的兴起，人文领域学者在参与数字人文研究过程中对于文化遗产数据资源的利用需求日益突显。多源、异构文化遗产信息资源的语义融合与互操作成为当前数字人文数据基础设施建设中的关键问题，而行之有效的实体语义相似度计算方法则成为实现这一目标的重要手段。[方法/过程] 以敦煌壁画叙词表关联数据为例，在分析该数据集本体模型与数据框架的基础上，针对其内容分布与结构特征提出一种多粒度匹配与加权运算相结合的实体语义相似度计算方法，并选取敦煌壁画叙词表关联数据中"飞天"相关实体为实验对象，引入属性特征、编辑距离等多种现有实体语义相似度计算方法进行对比实验。[结果/结论] 实验结果表明，本文提出的基于多粒度匹配的实体语义相似度计算方法，能够更好地适应敦煌壁画叙词表关联数据的内容与结构特征，在计算结果准确性方面比同类方法具有更好的表现，是推动数字人文背景下异构人文信息资源的数据互联与知识共享的又一可行思路。

关键词：敦煌壁画　关联数据　多粒度　语义相似度　实体相似度
分类号：G254　TP391
DOI：10.13266/j.issn.0252-3116.2021.08.010

1 引言

随着大数据、机器学习等技术的长足发展，以及多项文化遗产数字化典型实践的成功，数字人文与人文计算(Digital Humanities and Humanities Computing)已成为文化遗产资源组织领域中的新兴研究主题，受到学界、业界的广泛关注。数字人文与人文计算为新技术条件下文化遗产的数字化保护研究引入了新的思维模式，也丰富了传统人文学者利用文化遗产数据资源开展研究的应用场景。与此同时，强调在人文研究中力求过程可重复、数据可验证、方法可复用、结论可推广的数字人文研究范式也对文化遗产数据资源的整合、建构与组织质量提出了更高的要求。数据是人文计算的基石，文化遗产资源数据集的质量、颗粒度与覆盖范围等因素很大程度上决定了依托其开展的数字人文研究的成败、深度与可信度[1]。多源性、异构性是人文领域数据资源的典型特征，因此非结构化数据向结构化数据的转化成为数字人文在数据基础设施建设中的重要内容，在实践中这一过程主要通过关联开放数据(Linked Open Data)的构建与发布来实现，相关典型案例包括 Europeana、MuseumFinland、中国家谱关联数据、敦煌壁画叙词表关联数据等。

截至目前，国内外以关联数据支撑技术的文化遗产信息资源整合研究已经取得阶段性进展，以博物馆、美术馆、档案馆为代表的各类文化遗产保存服务机构依托实体馆藏开展数字化建设，在线发布了大量的文化遗产数据资源，为相关领域研究者提供了丰富的原始资料，极大地完善了数字人文研究的数据基础设施，满足了人文学者在参与数字人文研究过程中对基础数据的需求。而在初步解决了数据的来源问题后，数字人文下一步的数据基础设施建构应当走向高质量、宽领域与细粒度。在相关研究日益深入的背景下，推动文化遗产领域多源、异构数据集的聚合与融通已成为人文信息资源服务走向语义化、知识化、智能化的必要环节，而行之有效的语义相似度计算方法正是完成这一任务的关键技术之一。本文面向文化遗产领域数据

* 本文系国家社会科学基金重大项目"新时代我国文献信息资源保障体系重构研究"(项目编号:19ZDA345)与中央高校基本科研业务费自由探索项目"面向用户的文物信息资源知识服务研究"(项目编号:CCNU20A06025)研究成果之一。
作者简介：高劲松(ORCID:0000-0003-0022-5923)，教授，博士生导师，E-mail：jsgao@mail.ccnu.edu.cn；付家炜(ORCID:0000-0002-2996-3762)，博士研究生；李珂(ORCID:0000-0002-5212-9733)，硕士。
收稿日期:2020-09-28　修回日期:2021-01-08　本文起止页码:97-106　本文责任编辑:杜杏叶

资源的语义融合与互操作需求,提出了一种基于多粒度匹配的实体语义相似度计算方法,并以"敦煌壁画叙词表关联数据[2]"中"飞天"相关实体的语义相似度计算为例,探讨该方法应用于新阶段数字人文基础设施融合建构的价值与前景。

2 相关研究现状

数字人文与人文计算范式的快速兴起显著提升了人文领域研究者对高质量、宽领域、细粒度数据基础设施的需求。现有研究成果表明,以本体[3]、关联数据[4]、知识图谱[5]等为代表的语义网技术,在非结构化文化遗产资源向结构化语义数据集的转化过程中发挥了重要作用,能够面向数字人文研究的数据利用需求,有效支撑多种主题[6]、类型[7]、模态[8]和非结构化文化遗产数据资源的结构化整合。如何在依托上述模式构建的文化遗产语义数据集基础上,实现领域更宽、粒度更细、质量更优的多源融合与数据互操作,则成为新阶段数字人文数据基础设施建构中需要关注的重点问题。因此,构建行之有效的数据集实体语义相似度计算方法则成为实现这一目标的关键。

实体语义相似度计算的实质是通过求得具体数值,对一组命名实体间的相似性关系进行量化。近年来,随着关联开放数据标准下语义数据集创建与发布实践的日益丰富,国内外基于语义相似度的数据集语义关联发现研究也越来越多[9],其中产出了一系列不同的实体语义相似度计算策略,例如基于领域本体、语料库等依托外部数据的相似度算法[10],关联可视化、关联规则挖掘等基于内部数据驱动的相似度计算方法,以及基于路径、基于属性、基于内容等侧重于数据集体系特征的相似度计算方法。通过对上述实体语义相似度计算方法进行横向比较,可发现各类算法的主要差异体现在对语义数据集度量粒度的选择方面。粒度是用于比较数据、信息或知识粗糙性的度量单位,其精细度取决于数据集细化层次的深浅或划分模式的规模:层次越深、模式越多则粒度越细,反之则粒度更粗[11]。基于多粒度思想对主流的数据集实体语义相似度算法进行分类,其中粗粒度方法主要包括旭日图、树状图、圆堆图等数据可视化工具,以及多种基于路径距离的实体相似度算法;中粒度实体相似度计算方法则主要包括基于本体的属性特征分析、基于链接谓词的关联规则发现等;细粒度层面的实体相似度算法则主要通过挖掘实体的领域背景知识或上下文信息,以实现数据集中实体相关性的量化。上述实体语义相似度算法由于在度量粒度方面具有差异,因此各自的适用范围也各不相同。

在面向关联数据集的实体语义相似度计算研究中,路径距离是具有代表性的粗粒度实体语义相似度算法,此类方法将关联数据集的 RDF 三元组视为一种经典有向图模型,通过度量一组节点的路径距离反映其对应命名实体的语义相似度,Passant 方法[12]、Hickson 方法[13]均是将路径距离应用于数据集实体语义相似度计算的典型案例。属性特征则是应用较为广泛的中粒度实体语义相似度计算方法:语义关联数据的构建与发布往往伴随着与之对应的领域本体概念模型构建或复用,因而通过对本体模型中的类间关系与属性特征进行分析,能够有效支撑实体间的语义相关性的量化。Tversky 模型[14]是以本体属性特征判断实体语义相似度的典型算法,该模型主要依据一组实体间共有属性和差异属性的数量,对其语义相似度进行计算。路径距离与属性特征在语义数据集的实体语义相似度计算中各具优势,实践中往往将二者结合运用[15]:路径距离充分利用了 RDF 模型的三元组数据结构,在由节点与关系构成的关联语义网络中具有较高的运算效率和广泛的兼容性,但其在运算过程中将数据集中所有实体均视为无显著差异的节点,一定程度上忽略了其在细粒度层面的特征关系,在面向多个复杂数据集的实体语义相似度计算中,存在误差大、开销大等问题。此外,对于单一的路径距离算法,也难以适用于跨领域数据集的实体相似度计算需求。而通过与属性特征方法的结合,路径距离算法在细粒度层面的缺陷将得到较好补足,同时也规避了基于本体的语义相似度算法对于数据集构建质量与构建方式的较高限制[16]。

随着语义关联数据相关实践的不断深入,大规模知识库中包含的实体规模快速增长,同时实体的属性特征和标注层次也不断细化。面向语义数据集的实体语义相似度计算方法设计日益呈现多粒度、多方法融合的特征。例如:贾丽梅等[17]在关联数据属性特征的基础上,通过引入基于动态权值的语义相似度算法和面向属性重要性与取值类型的动态加权机制,提升了语义相似度计算的准确性。R. Meymandpour 等[18]提出了一种基于上下文的关联数据相似度计算策略,通过 SPARQL 查询全面获取关联数据集的属性列表及各项属性的取值内容,并引入基于语料库的词向量模型进行语义相似度的计算。刘晓娟等[19]基于对关联数据的隐含知识网络特性的分析,提出了一种改进的向量空间模型,并通过引入属性加权思想进一步提升了关

联数据实体语义相似度的计算精度。上述研究也反映出，现阶段面向数据集的实体语义相似度计算方法，在设计思路上逐步从算法技术导向转换为对象需求导向，在方法设计过程中更加注重对实体所在数据集领域背景知识和模型框架结构的分析，并通过加权运算方式对面向不同粒度的实体相似度计算方法进行整合[20]，从而在语义数据集构建技术快速演进更迭的背景下，进一步保证并提升实体语义相似度计算结果的准确性与可靠性。通过分析上述研究现状，可以看出目前面向数字人文与人文计算的文化遗产资源的语义组织研究已经取得了初步成果，国内外研究者通过对本体、关联数据、知识图谱等语义组织工具的综合应用，着眼于不同的细分领域文化遗产信息资源的内容与形式特征，对文化遗产信息资源的开发、利用共享进行了卓有价值的探索。为了更好地满足人文领域研究者参与数字人文研究过程中对于高质量、宽领域、细粒度文化遗产数据的利用需求，有必要进一步推动语义网技术与文化遗产资源组织的融合。通过深入研究文化遗产领域信息资源的概念集成、本体匹配和实体关联发现方法，提升相关领域中数据集成和知识价值开发复用的效能。本文着眼于数字人文背景下文化遗产语义数据集的实体关联发现问题，以敦煌壁画叙词表关联数据为例，探讨如何在有机整合现有的实体语义相似度计算方法基础上，对文化遗产关联数据的实体语义相关性进行有效量化，进而为多源异构文化遗产数据资源的语义融合与数据互操作提供可行思路。

3 敦煌壁画叙词表关联数据的实体语义相似度计算方法

3.1 敦煌壁画叙词表关联数据的基本概况

敦煌学是中国文化遗产研究中的一个特殊领域，敦煌壁画更是人类文化遗产中的瑰宝，具有极高的艺术观赏和科学研究价值。随着文化遗产数字化和数字人文研究的兴起，敦煌研究者积累了大量的一手信息资源，为敦煌学研究和敦煌壁画的传播提供了重要条件。为了发掘敦煌壁画资源中蕴含的语义信息，并对其进行有效组织和规范表达，国内学者围绕敦煌壁画数字资源的语义标注、信息检索与知识组织需求，在对以 AAT（艺术与建筑叙词表）为代表的多层级结构化叙词表进行调研分析的基础上，整合《敦煌学大辞典》《敦煌石窟内容总录》《敦煌人物志》等敦煌学基础文献，通过自顶向下与自底向上相结合的构建方法完成了敦煌壁画叙词表的编制，并利用语义网技术实现叙词表的关联数据发布[21]。该研究的核心成果“敦煌壁画叙词表关联数据”已成为当前文化遗产语义组织领域具有代表性的实践案例之一，已发布的敦煌壁画叙词表关联数据集含有语义实体 4 500 余个，三元组规模达 27 500 余条，涉及敦煌壁画叙词表的 5 大分面，25 个二级类目，3 896 个受控词汇，能够为敦煌壁画数字资源的深度语义标注、语义检索、知识组织、信息关联与共享等提供有效的数据支撑[22]。

3.2 敦煌壁画叙词表关联数据的语义描述粒度分析

关联数据通过 RDF 三元组实现资源的描述与组织，在三元组中由链接谓词（Predicate）在头部实体（Subject）和尾部实体（Object）之间建立链接，以描述不同资源之间存在的属性关联关系。在关联数据发布实践中，用于描述特定资源的语义实体往往由多条三元组共同构成，由于链接谓词的不同，实体中各个三元组的语义描述粒度往往存在差异。在语义相似度计算过程中，实体之间的层次关系、逻辑关系和属性参数对于相似度计算结果均具有不同程度的影响，如果采用单一的计算方法对粒度不同的多种三元组进行直接比较，往往会造成语义信息的丢失，进而产生计算误差。因此在敦煌壁画叙词表关联数据的实体语义相似度计算过程中，有必要通过分析链接谓词的构成来揭示不同类型三元组的语义描述粒度，在此基础上为不同粒度层级的语义描述模块匹配相适应的语义相似度计算方法。

本体构建是关联数据创建与发布的重要环节，本体模型通过定义类与类的属性关系以描述资源实体之间的语义关系和层级结构。敦煌壁画叙词表本体[23]是在敦煌壁画叙词表逻辑结构基础上，通过复用 GVP 本体、SKOS 数据模型和 DCMI 元数据标准中的术语元素构建的本体模型。该本体定义了敦煌壁画叙词表关联数据的层级结构，为敦煌壁画叙词表的语义转化和关联数据发布提供了术语框架。在敦煌壁画叙词表关联数据实体语义相似度计算过程中，通过分析敦煌壁画叙词表本体的 Schema 框架，能够对敦煌壁画关联数据中的链接谓词进行全面抽取，进而在此基础上有效揭示其语义描述粒度。敦煌壁画叙词表本体的属性定义见表 1，根据描述功能的不同分为对象属性和数据属性。对象属性用于描述类与类之间的相关关系，大多数对象属性仅用于描述一组概念之间的相关关系，如 exactMatch、related 属性用于描述概念间的相同或相关关系，inScheme、hasTopConcept 属性用于描述概念与

词表、词表与分面之间的包含关系。而 broader、narrower 属性则被定义为多种类间关系的描述媒介，其既能够描述概念之间的上下位关系，也能够描述概念与分面(Facet)之间的层级关系。数据属性则用于描述实体在不同方面的性质，例如名称、创建时间、创建者等具体信息，其属性值的数据类型多为短文本字符型，仅有 scopeNote 属性因专用于著录抽取自领域专业文献的背景知识，其属性值类型为长文本型。

表1 敦煌壁画叙词表本体的属性定义

属性分类	属性名	定义域(domain class)	值域(range class)
对象属性	skos:broader	skos:Concept	skos:Concept、gvp:Facet、gvp:Hierarchy
	skos:narrower	skos:Concept、gvp:Facet	skos:Concept、gvp:Hierarchy
	skos:exactMatch	skos:Concept	gvp:Concept
	skos:hasTopConcept	skos:ConceptScheme	gvp:Facet
	skos:inScheme	skos:Concept	skos:ConceptScheme
	skos:related	skos:Concept	skos:Concept
	rdf:type	skos:Concept	skos:Concept、dhvocab:instance
数据属性	skos:preLabel	skos:Concept	<value>(概念名称)
	skos:scopeNote	skos:Concept	<value>(概念背景知识)
	dc:created	skos:Concept	<value>(概念创建时间)
	dc:creator	skos:ConceptScheme	<value>(词表创建者)
	dc:rights	skos:ConceptScheme	<value>(词表版权所有者)
	dc:title	skos:ConceptScheme	<value>(词表正式题名)

综上所述，基于对敦煌壁画叙词表本体模型和 Schema 框架的分析，本文将敦煌壁画叙词表关联数据的语义描述粒度划分为以下三个层次：①粗粒度层级，由反映敦煌壁画叙词表关联数据中不同实体层级结构关系的三元组构成，对应的链接谓词包括反映概念上下位关系的 broader、narrower、hasTopConcept 属性以及反映概念共指关系的 exactMatch 属性。②中粒度层级，由反映叙词表关联数据中实体之间逻辑关系信息的三元组构成，对应的链接谓词包括反映实体语义关系的 type、inScheme、related 等对象属性以及反映实体固有性质的 preLabel、created、creator、rights 等短文本属性。③细粒度层次，由标注叙词表中部分实体所具有领域背景信息的三元组构成，对应的链接谓词为长文本属性 scopeNote。

3.3 基于多粒度匹配的实体语义相似度计算模型

现阶段的实体语义相似度计算方法研究逐渐从单一的技术导向转换为面向计算对象特征的需求导向，更加注重对实体所在数据集领域的背景知识和模型框架结构的分析，面向不同粒度的实体三元组匹配与之适应的方法，来进行语义相似度计算。本节在上述思想基础上，依据对敦煌壁画叙词表关联数据语义描述粒度的分析结果，提出一种多粒度匹配与加权运算相结合的实体语义相似度计算模型，其基本框架见图1。首先，通过敦煌壁画叙词表关联数据的 SPARQL 查询端口访问并获取待计算实体的三元组数据，包括头部实体、尾部实体与链接谓词；其次，根据三元组中链接谓词对应的语义粒度层级将其与模型中的粗、中、细粒度模块进行匹配；再次，针对各模块三元组的内容与结构特点分别设置与之对应的计算方法并完成语义相似度的计算；最后，依据各模块三元组中链接谓词的构成情况进行权重分配，并在此基础上通过加权运算得出该组实体的综合语义相似度。

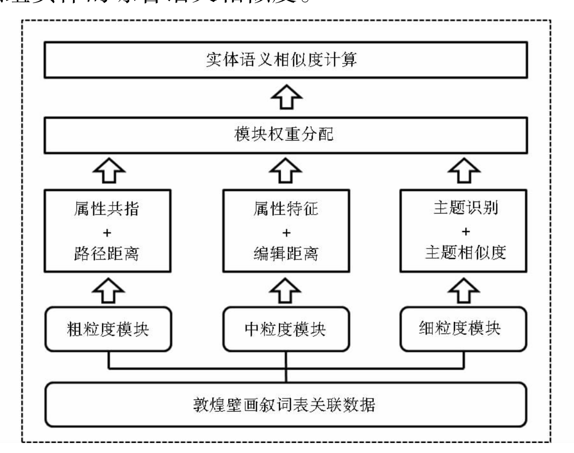


图1 基于多粒度匹配的实体语义相似度计算模型

3.3.1 粗粒度模块的实体相似度计算方法

粗粒度模块面向敦煌壁画叙词表关联数据中用于描述实体层级结构关系的三元组，模型采用属性共指与路径距离相结合的语义相似度计算方法。

(1)基于属性共指的语义相似度计算。在对敦煌壁画叙词表关联数据中两个实体进行语义相似度计算

前,首先应判断二者间是否具有等价属性。如果两个实体之间存在如 owl:sameAS、rdfs:seeAlso 或 skos:exactMatch 等表示共指关系的层次属性,其语义相似度应判断为 1,否则这一部分的相似度为 0,其计算公式如公式 1 所示:

$$Sim_{same}(x,y)=\begin{cases}1, & sameAs\\ 0, & otherwise\end{cases} \quad 公式(1)$$

(2)基于路径距离的语义相似度计算。作为敦煌壁画叙词表的语义发布成果,实体之间的层级性是敦煌壁画叙词表关联数据的重要特性。因此在语义相似度计算过程中,应当充分考虑各个实体之间的层级关系特征,引入面向概念相对深度的语义相似度计算思想[24]。路径距离即为遵循这一思想的语义相似度计算方法:两个实体之间的路径距离越短,则其语义相似度越高,其计算公式见公式 2[25]。其中 length(x,y)表示实体 x、y 在概念层次结构树中的路径长度(即从 x 链接到 y 的跳转次数),α 为调节参数,通常可以取值为 1。

$$Sim_{Route_Dis}(x,y)=\frac{\alpha}{min[\,length(x,y)\,]+\alpha} \quad 公式(2)$$

3.3.2 中粒度模块的实体相似度计算方法

中粒度模块面向敦煌壁画叙词表关联数据中用于描述实体固有属性及相关关系的三元组,模型采用属性特征与编辑距离相结合的语义相似度计算方法。

(1)基于属性特征的语义相似度计算。在关联数据中以对象属性为链接谓词的三元组能够描述头部实体与尾部实体间存在的特定语义关系。因此不同实体之间所含对象属性的异同情况能够有效反映其语义相关程度。Tversky 模型是基于属性特征计算实体语义相似度的典型方法,该模型依据一对实体含有的公共属性与差异属性的数量,利用公式 3 所示的运算方法对二者语义相似度进行量化[14]。其中 $f(x\cap y)$ 表示实体 X、Y 含有的公共属性的数量,$f(x-y)$ 表示实体 x 包含而实体 y 不包含的属性数量,反之 $f(y-x)$ 则表示实体 y 包含而实体 x 不包含的属性数量。α、β 为调节参数,用于反映实体 X、Y 的重要程度,默认取值为 1。

$$Sim_{Tversky}(x,y)=\frac{f(x\cap y)}{f(x\cap y)+\alpha f(x-y)+\beta f(y-x)} \quad 公式(3)$$

在引入 Tversky 模型的基础上,还需结合敦煌壁画叙词表关联数据的具体特性对其进行必要改进。一对实体虽然具有某项公共属性,但是该属性在各自三元组中对应的宾语实体却不尽相同。针对这一现象,本模型在公式 3 基础上进行如下调整:对于一组链接谓词(Predicate)相同的属性,只有其在三元组中链接的尾部实体(Object)也相同时,才将其视为两个实体的公共属性,否则均视为所在实体的独有属性,在公式 3 中记入分母部分。

(2)基于编辑距离的语义相似度计算。编辑距离是语义相似度计算的典型方法,在本模型中主要用于计算 skos:prefLabel、dc:created 等短文本属性值的语义相似度。该方法采用转化思想对原始实体和目标实体的属性值文本相似度进行量化[26],计算公式见公式 4。其中 tc(x – y)表示 x 向 y 转换所需的最小次数,操作内容包括属性值的加减、插入、替换和删除等,max[|x|, |y|]表示两个属性值的最大字长。

$$Sim_{EDIT_Dis}(x,y)=1-\frac{tc(x\text{-}y)}{max[\,|x|,|y|\,]} \quad 公式(4)$$

3.3.3 细粒度模块的实体相似度计算方法

细粒度模块面向敦煌壁画叙词表关联数据中用于著录实体相关背景信息的三元组,主要针对长文本属性 skos:scopeNote 的值进行语义相似度计算。由于长文本属性值往往包含多个语句段落,文本结构复杂且信息容量较高,因此上文针对短文本属性值的编辑距离方法往往难以适用。面向长文本信息的语义相似度计算需求,本模型采用主题识别与 Tversky 模型相结合的主题相似度计算策略,首先使用文本主题识别工具,从原始实体和目标实体的长文本属性值中分别抽取规定数量的主题词。再统计二者共有主题词和独有主题词的数量,并代入 Tversky 模型以量化其语义相似度,计算过程如公式 5 所示:

$$Sim_{scopeNote}(x,y)=\frac{Count(x\cap y)}{Count(x\cap y)+Count(x-y)+Count(y-x)} \quad 公式(5)$$

上文针对敦煌壁画叙词表关联数据中粗粒度、中粒度与细粒度模块的实体语义相似度分别提出了相应的计算方法。在实际计算过程中,还需通过分析计算对象的内容分布、属性特征等具体情形,合理设定三个粒度模块中各个计算方法的权重系数,以得出该组实体的综合语义相似度,计算过程如公式 6 所示(α、β、γ 为各个模块权重系数):

$$Sim_{total}=\alpha Sim_{粗粒度}+\beta Sim_{中粒度}+\gamma Sim_{细粒度} \quad 公式(6)$$

4 敦煌壁画叙词表关联数据实体语义相似度计算实验

4.1 数据来源

为了验证上文提出的计算方法在敦煌壁画叙词表关联数据实体语义相似度计算中的实际效果，本文以数据集中“飞天”相关实体为实验对象，引入多种同类算法开展语义相似度计算的对比实验。实验数据通过 SPARQL 查询方式从敦煌壁画叙词表关联数据服务平台获取：在平台 SPARQL Endpoint 端口[27]中使用图 2 所示的 SPARQL 查询式对 skos：prefLabel 属性值中含有“飞天”字段的所有实体进行检索，共获取有效实体 8 个，分别为：①双飞天 < dhvocab：tema2875 >；②飞天髻 < dhvocab：tema3655 >；③莲花飞天藻井图案 < dhvocab：tema1993 >；④飞天纹 < dhvocab：tema445 >；⑤飞天 < dhvocab：tema245 >；⑥飞天乐伎 < dhvocab：tema2533 >；⑦中原式飞天 < dhvocab：tema2551 >；⑧西域式飞天 < dhvocab：tema2552 >。查询结果见图 3。



图 2 SPARQL 查询式

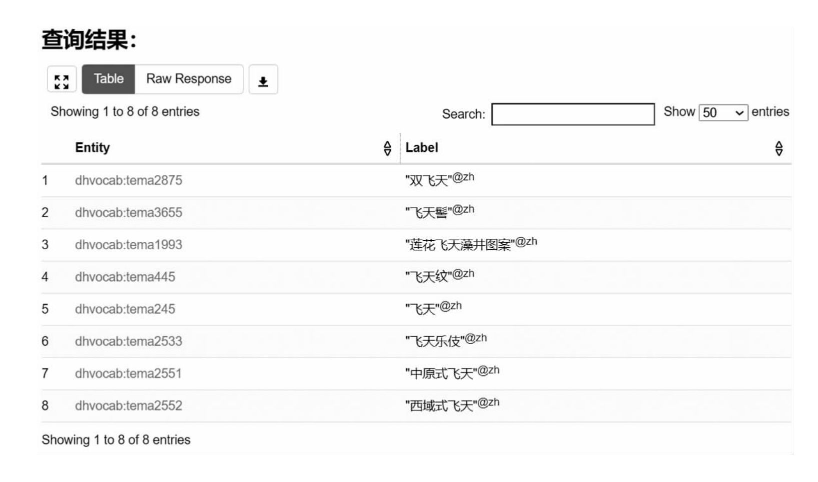

查询结果：

Table | Raw Response

Showing 1 to 8 of 8 entries　　Search:　　Show 50 entries

	Entity	Label
1	dhvocab:tema2875	"双飞天"@zh
2	dhvocab:tema3655	"飞天髻"@zh
3	dhvocab:tema1993	"莲花飞天藻井图案"@zh
4	dhvocab:tema445	"飞天纹"@zh
5	dhvocab:tema245	"飞天"@zh
6	dhvocab:tema2533	"飞天乐伎"@zh
7	dhvocab:tema2551	"中原式飞天"@zh
8	dhvocab:tema2552	"西域式飞天"@zh

Showing 1 to 8 of 8 entries

图 3 SPARQL 查询结果

4.2 实验过程

4.2.1 实验内容

将上文获取的 8 条关联数据实体两两分组，构建 8 ×8 的实体相似度矩阵，生成 28 条语义相似度计算任务，见表 2。下文分别使用基于多粒度匹配的实体相似度、基于 Tversky 模型的属性特征相似度和基于编辑距离的标签文本相似度方法进行实体语义相似度的计算实验。

表 2 飞天相关实体语义相似度矩阵

实体	①tema245	②tema445	③tema1993	④tema2533	⑤tema2551	⑥tema2552	⑦tema2875	⑧tema3655
①tema245	—	T1：①②	T2：①③	T3：①④	T4：①⑤	T5：①⑥	T6：①⑦	T7：①⑧
②tema445	—	—	T8：②③	T9：②④	T10：②⑤	T11：②⑥	T12：②⑦	T13：②⑧
③tema1993	—	—	—	T14：③④	T15：③⑤	T16：③⑥	T17：③⑦	T18：③⑧
④tema2533	—	—	—	—	T19：④⑤	T20：④⑥	T21：④⑦	T22：④⑧
⑤tema2551	—	—	—	—	—	T23：⑤⑥	T24：⑤⑦	T25：⑤⑧
⑥tema2552	—	—	—	—	—	—	T26：⑥⑦	T27：⑥⑧
⑦tema2875	—	—	—	—	—	—	—	T28：⑦⑧
⑧tema3655	—	—	—	—	—	—	—	—

4.2.2 基于多粒度匹配的实体语义相似度计算

使用基于多粒度匹配的方法进行实体语义相似度计算需要事先对待计算实体的三元组构成情况进行分析，根据不同粒度层级中三元组的分布特征为各个粒

度模块分配相应的权重系数。笔者通过敦煌壁画叙词表关联数据服务平台提供的数据获取端口下载上文所述8个实体的RDF文档，对其包含的55个链接谓词进行分类并分别统计各类属性的占比，统计结果如表3所示：

表3 链接谓词数量占比统计

属性类型	链接谓词	数量	占比
层级属性	skos:broader;skos:narrower	20	36.36%
对象属性	rdf:type;skos:inScheme	16	29.09%
短文本属性	dct:created;skos:prefLabel	16	29.09%
长文本属性	skos:scopeNote	3	5.45%

(1)粗粒度模块中，由于实验涉及的8个实体均不包含 owl:sameAs、rdfs:seeAlso、skos:exactMatch 等共指属性，因此本次实验中可不考虑实体等价对语义相似度的影响。如表3所示，8个实体中共包含20个反映层级属性的链接谓词，在所有属性中占比最高，可知在本数据集中实体间的路径距离对其语义相似度计算的影响程度较高，依据链接谓词占比将粗粒度模块的权重系数定义为0.363 6。

(2)中粒度模块中，8个实体中共含反映对象属性的链接谓词16项，依据其占比将本模块中改进 Tversky 模型算法的权值设为0.290 9。此外，反映短文本属性的链接谓词数量也为16项，因此本模块中编辑距离相似度算法的权值亦设为0.290 9。其中由 dc:created 属性标注的日期数据需转化为时间戳文本后再进行编辑距离计算。

(3)细粒度模块中，由于8个实体中仅飞天 < dhvocab:tema245 >、莲花飞天藻井图案 < dhvocab:tema1993 >、双飞天 < dhvocab:tema2875 >实体中各含有1项 skos:scopeNote 属性，可见在本实验中各实体的领域背景信息对于计算结果的影响较小，故依据其占比将细粒度模块权重系数定为0.054 5。

综上所述，在分析各模块链接谓词构成情况的基础上定义如表4所示的实验方案，并据此分别完成粗粒度、中粒度与细粒度模块的语义相似度计算。

表4 基于多粒度匹配的飞天实体语义相似度计算

粒度层级	计算对象	计算方法	权重系数
粗粒度模块	层级属性	路径距离	0.363 6
中粒度模块	对象属性	改进 Tversky 模型	0.290 9
	短文本属性	文本编辑距离	0.290 9
细粒度模块	长文本属性	主题相似度	0.054 5

此处以任务"T1:Sim(tema245,tema445)"为例，阐述基于多粒度匹配的实体语义相似度计算过程：在粗粒度模块中，飞天 < dhvocab:tema245 > 与飞天纹 < dhvocab:tema445 > 的路径距离为10，代入公式2可知其路径距离相似度为0.090 9，加权后为0.033 1。在中粒度模块中，经改进 Tversky 模型计算，飞天 < dhvocab:tema245 > 与飞天纹 < dhvocab:tema445 > 的属性特征相似度为0.4，加权后为0.1164；短文本属性经编辑距离(公式4)计算得到相似度0.633 3，加权后为0.184 2。在细粒度模块中，由于飞天纹 < dhvocab:tema445 > 中不包含 skos:scopeNote 属性，因此二者细粒度模块的语义相似度为0。综上，飞天 < dhvocab:tema245 > 与飞天纹 < dhvocab:tema445 > 的语义相似度为0.333 6。采用相同方法可计算其他27组实体的相似度。

4.2.3 基于 Tversky 模型的属性特征相似度计算

如公式3所示，经典 Tversky 模型在计算两个实体语义相似度的过程中仅依据二者共有属性和差异属性的数量进行计算，而不考虑属性的具体取值情况。例如在任务 T1 中，实体飞天 < dhvocab:tema245 > 与飞天纹 < dhvocab:tema445 > 中属性相同的三元组为5项，飞天 < dhvocab:tema245 > 含有独有属性2项，飞天纹 < dhvocab:tema445 > 不含独有属性，则代入公式3可知其相似度为0.714 3。采用相同算法即可完成其他27项相似度计算任务。

4.2.4 基于编辑距离的标签文本相似度计算

基于编辑距离的文本相似度计算是大规模关联数据融合与互操作实践中的常用方法，其基本思想是：关联数据实体中 dc:title、skos:prefLabel 等用于反映题名、标签信息的属性，其取值均为数据创建或发布者从自然语言中精选而来的具有代表性、规范性的语词，基于标签编辑距离进行语义相似度计算能够较好地平衡计算效率、结果质量和性能开销。标签编辑距离采用如公式4所示基于转化的计算思想：通过一组实体属性值最短编辑次数与最大字长的比值衡量其语义相似度的高低。此处仍以 T1 为例：实体 < dhvocab:tema245 > 与 < dhvocab:tema445 > 的 skos:prefLabel 属性值分别为"飞天"和"飞天纹"，其最短编辑距离为1，最大字长为3，代入公式4可知二者的标签编辑距离相似度为0.666 7。其他27项相似度计算任务亦采用相同算法完成。

4.3 实验分析

分别使用基于经典 Tversky 模型、标签编辑距离和多粒度匹配的实体语义相似度计算方法完成表1中的28个计算任务，结果如表4所示：

表 4 实验结果

计算任务	计算对象	Tversky 模型	标签编辑距离	多粒度匹配	计算任务	计算对象	Tversky 模型	标签编辑距离	多粒度匹配
T01	tema245;tema445	0.714 3	0.666 7	0.333 6	T15	tema1993;tema2551	0.833 3	0.125 0	0.247 8
T02	tema245;tema1993	0.857 1	0.250 0	0.280 8	T16	tema1993;tema2552	0.833 3	0.125 0	0.247 8
T03	tema245;tema2533	0.714 3	0.500 0	0.300 6	T17	tema1993;tema2875	1.000 0	0.250 0	0.328 5
T04	tema245;tema2551	0.714 3	0.400 0	0.395 1	T18	tema1993;tema3655	0.833 3	0.250 0	0.209 8
T05	tema245;tema2552	0.714 3	0.400 0	0.395 1	T19	tema2533;tema2551	1.000 0	0.000 0	0.351 5
T06	tema245;tema2875	0.857 1	0.666 7	0.338 7	T20	tema2533;tema2552	1.000 0	0.000 0	0.351 5
T07	tema245;tema3655	0.714 3	0.666 7	0.231 6	T21	tema2533;tema2875	0.833 3	0.250 0	0.295 1
T08	tema445;tema1993	0.833 3	0.250 0	0.271 5	T22	tema2533;tema3655	1.000 0	0.500 0	0.242 4
T09	tema445;tema2533	1.000 0	0.500 0	0.302 3	T23	tema2551;tema2552	1.000 0	0.400 0	0.615 7
T10	tema445;tema2551	1.000 0	0.200 0	0.260 7	T24	tema2551;tema2875	0.833 3	0.400 0	0.318 9
T11	tema445;tema2552	1.000 0	0.200 0	0.260 7	T25	tema2551;tema3655	1.000 0	0.200 0	0.215 1
T12	tema445;tema2875	0.833 3	0.333 3	0.256 1	T26	tema2552;tema2875	0.833 3	0.400 0	0.318 9
T13	tema445;tema3655	1.000 0	0.666 7	0.243 6	T27	tema2552;tema3655	1.000 0	0.200 0	0.215 1
T14	tema1993;tema2533	0.833 3	0.250 0	0.264 2	T28	tema2875;tema3655	1.000 0	0.333 3	0.230 3

此处以表 3 中 T4、T7、T23 的计算结果(见图 4)为例,比较三种算法在"飞天"相关实体语义相似度计算中的效果。上述 3 个任务的基本概况如下:

(1)T4:Sim(tema245,tema2551)的计算对象为实体"飞天 < tema245 >"和"中原式飞天 < tema2551 >",在"敦煌壁画叙词表关联数据"中,前者是后者的上位概念(< dhvocab:tema2551 > < skos:broader > < dhvocab:tema245 >),二者的路径距离为 1,具有较高的语义相关度。通过比较不同方法在 T4 中的语义相似度计算结果,能够凸现各方法对于实体路径距离的敏感程度。

(2)T7:Sim(tema245,tema3655)的计算对象为实体"飞天 < tema245 >"和"飞天髻 < tema3655 >",在"敦煌壁画叙词表关联数据"中,前者是实体"佛家神祇 < tema204 >"的下位概念(< dhvocab:tema204 > < skos:narrower > < dhvocab:tema245 >),是对一类特定佛教人物的统称;后者是实体"发式 < tema3640 >"的下位概念(< dhvocab:tema3655 > < dhvocab:instance > < dhvocab:tema3640 >),用于描述壁画人物的一种造型风格。二者的标签文本虽然相似,但在数据集之中的路径距离高达 12,实际的语义相关度也较低,通过比较不同方法在 T7 中的计算结果,能够直观判断各方法对于标签内容相似但语义关联较低的"易错"实体的识别效果。

(3)T23:Sim(tema2551,tema2552)的计算对象为实体"中原式飞天 < tema2551 >"和"西域式飞天 < tema2552 >",在"敦煌壁画叙词表关联数据"中,二者均为实体"飞天 < tema245 >"的下位概念,用于描述"飞天"意象在不同地域文化中的形象风格。在数据集之中,二者的路径距离(距离为 2)虽然大于任务 T4 中两实体的路径距离(距离为 1),但在先验知识层面具有更高的语义相似度,因此 T23 适合用于比较不同计算方法对于此类隐性高相关度实体的识别能力。

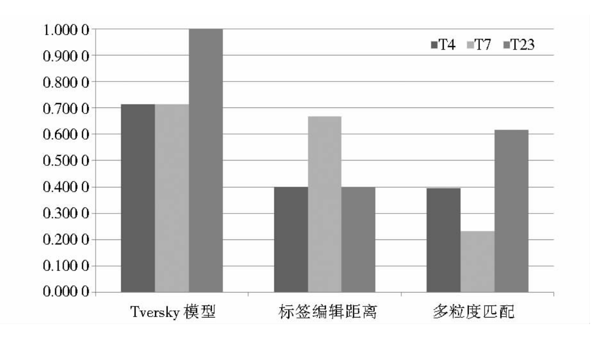


图 4 实体语义相似度计算结果对比

首先,基于经典 Tversky 模型的计算结果为:T4:Sim(tema245,tema2551) = T7:Sim(tema245,tema3655) < T23:Sim(tema2551,tema2552)。可以看出,相比其他两种方法,基于 Tversky 模型的语义相似度计算结果整体偏高。其原因在于:由于领域范畴和标注对象的内容结构特点,敦煌壁画叙词表关联数据中各个实体普遍呈现属性数量较少且重复程度较高的基本特性。对于面向属性特征的经典 Tversky 模型而言,上述特性易导致其语义相似度计算结果出现数值偏高且区分度不足的问题。因此,在应用 Tversky 模型进行关联数据语义相似度计算的实践中,有必要依据计算需求对其进行必要的改进,通过调整实体间共有属性的判别标准,以规避上述现象对于计算结果的干扰。

其次,基于标签编辑距离的计算结果为:T4:Sim

(tema245,tema3655) = T23:Sim(tema2551,tema2552) < T7:Sim(tema245,tema3655)。可以看出本方法对 T4 的计算结果相对准确,但在 T7、T23 的语义相似度计算中存在较大误差。其原因在于:该方法直接以实体标签文本内容作为语义相似度评判依据,对于文化遗产关联数据中诸如“飞天 < tema245 >”、“飞天髻 < tema3655 >”(字面相似但语义关联度低)以及“中原式飞天 < tema2551 >”、“西域式飞天 < tema2552 >”(字面不相似但语义关联度高)这类标签内容与语义距离不一致的特殊实体,往往难以准确计算其语义相似度。

再次,基于多粒度匹配的计算结果为:T7:Sim(tema245,tema3655) < T4:Sim(tema245,tema2551) < T23:Sim(tema2551,tema2552),与上文对 T4、T7、T23 的先验知识描述基本一致。其原因在于:基于多粒度匹配的计算方法能够根据数据集的内容结构特点,对其构成要素进行较为合理的粒度划分,并针对各个模块分别选取与之适应的具体计算方法;在面向领域背景知识丰富、层次结构复杂的敦煌壁画叙词表关联数据进行语义相似度计算的过程中,相比其他基于单一思路的计算方法能够取得准确性更优的计算结果。

通过对三种方法的计算结果进行比较,能够得出以下认识:在利用语义相似度对文化遗产领域的关联数据集进行语义融合与互操作的过程中,受制于相关领域的背景知识复杂、不同实体之间的语义边界模糊等客观条件的影响,有必要在对领域本体模型和关联数据 Schema 框架进行充分调研分析的基础之上选取针对性的计算方法。同时,文化遗产领域的知识结构多维性,也使得基于单一策略的计算方法难以全面满足数据集内所有实体的语义相似度计算需求。基于这一背景,面向文化遗产领域的关联数据语义相似度计算应当遵循以下思路:首先,应在语义描述粒度分析的基础上对关联数据集进行模块化处理;其次,应面向不同模块的内容与结构特征选取相适应的语义相似度计算方法,并在此基础上通过合理设置各个模块的权值系数以获取最优的语义相似度计算结果。

5 结语

本文面向人文计算研究范式兴起的背景下,人文学者参与数字人文研究过程中对文化遗产领域数据集的语义融合与互操作需求,以敦煌壁画叙词表关联数据为例,在数据集语义描述粒度分析的基础上提出了一种基于多粒度匹配的实体语义相似度计算方法,为数字人文背景下异构人文信息资源的数据互联与知识共享提供了一种可行思路。本文在分析敦煌壁画叙词表关联数据的本体模型和数据结构的基础上,依据数据集中实体间的层级关系、逻辑关系、属性参数等构成要素对数据集三元组的语义粒度层级进行划分。其次,针对数字人文领域多源异构数据集的知识融合需求,提出基于多粒度匹配的实体语义相似度计算模型,依据不同粒度下实体在数据集中的内容与结构特征,合理匹配与之适应的语义相似度算法,进而实现了计算需求与计算方法的有机整合。在实验部分,本文以敦煌壁画叙词表关联数据中的“飞天”相关实体为例,采用本文提出的多粒度匹配方法,与当前具有代表性的属性特征相似度、标签编辑距离相似度方法进行语义相似度计算对比实验。实验结果表明,本文提出的基于多粒度匹配的实体语义相似度计算方法能够更好地适应敦煌壁画叙词表关联数据领域背景知识复杂、实体语义边界模糊的结构特性,相比其他两种基于单一策略的语义相似度算法能够取得准确性更优的计算结果。在未来的研究中,还可进一步将本文提出的计算方法运用于文化遗产领域其他的关联数据集中,通过开展跨数据集的大规模语义相似度计算实验,对不同粒度下的权值分配、不同算法中的参数设置等技术细节进行调整与优化。

参考文献:

[1] 黄水清. 人文计算与数字人文:概念、问题、范式及关键环节[J]. 图书馆建设,2019(5):68 - 78.

[2] 敦煌壁画叙词表关联数据服务平台[EB/OL]. [2021 - 01 - 06]. http://dh.whu.edu.cn/dhvocab/home.

[3] 左丹,欧石燕. 人文信息资源语义描述、语义组织研究与实践述评[J]. 图书馆论坛,2019,39(8):21 - 31.

[4] 侯西龙,谈国新,庄文杰,等. 基于关联数据的非物质文化遗产知识管理研究[J]. 中国图书馆学报,2019,45(2):88 - 108.

[5] 陈涛,刘炜,单蓉蓉,等. 知识图谱在数字人文中的应用研究[J]. 中国图书馆学报,2019,45(6):34 - 49.

[6] 夏翠娟,张磊. 关联数据在家谱数字人文服务中的应用[J]. 图书馆杂志,2016,35(10):26 - 34.

[7] 翟姗姗,许鑫,夏立新,等. 语义出版技术在非遗数字资源共享中的应用研究[J]. 图书情报工作,2017,61(2):23 - 31.

[8] 曾子明,周知,蒋琳. 基于关联数据的数字人文视觉资源知识组织研究[J]. 情报资料工作,2018(6):6 - 12.

[9] 龚振,范冰冰. 数据集的语义关联发现方法研究[J]. 计算机应用与软件,2018,35(8):83 - 86,185.

[10] 张哲. 基于语义相似度分析的关联数据模型研究[D]. 北京:北京邮电大学,2018.

[11] 王忠义,周杰,黄京. 数字图书馆多粒度关联数据的创建与发布[J]. 情报学报,2016,35(8):885 - 896.

[12] PASSANT A. Measuring semantic distance on linking data and i-

sing it for resources recommendations[C]//Aaai spring symposium: linked data meets artificial intelligence. 2010:93 – 98.
[13] HICKSON M, KARGAKIS Y, TZITZIKAS Y. Similarity-based browsing over linked open data[EB/OL]. [2021 – 04 – 03]. https://arxiv.org/pdf/1106.4176v1.pdf.
[14] TVERSKY A. Features of similarity[J]. Readings in cognitive science,1977,84(4):290 – 302.
[15] 邓兰兰,李春旺. 关联数据资源集相似度计算方法研究[J]. 情报理论与实践,2012,35(5):112 – 116.
[16] 孙海霞,钱庆,成颖. 基于本体的语义相似度计算方法研究综述[J]. 现代图书情报技术,2010(1):51 – 56.
[17] 贾丽梅,郑志蕴,李钝,等. 基于动态权值的关联数据语义相似度算法研究[J]. 计算机科学,2014,41(8):263 – 266,273.
[18] MEYMANDPOUR R, DAVIS J. A semantic similarity measure for linked data: an information content-based approach[J]. Knowledge-based systems,2016,109(19):276 – 293.
[19] 刘晓娟,刘群. 基于关联数据的探索式检索系统研究与实现[J]. 图书情报工作,2017,61(5):117 – 124.
[20] 张立波,孙一涵,罗铁坚. 一种基于大规模知识库的语义相似性计算方法[J]. 计算机研究与发展,2017,54(11):2576 – 2585.
[21] 王晓光,侯西龙,程航航,等. 敦煌壁画叙词表构建与关联数据发布[J]. 中国图书馆学报,2020,46(4):69 – 84.
[22] 敦煌壁画叙词表项目介绍[EB/OL]. [2021 – 01 – 06]. http://dh.whu.edu.cn/dhvocab/dhresource/html/intro.html.
[23] 本体模型[EB/OL]. [2021 – 01 – 06]. http://dh.whu.edu.cn/dhvocab/ontology.
[24] RADA R, MILI H. Development and application of a metric on semantic nets[J]. Ieee transaction on system man & cybernetics, 1989, 19(1):17 – 30.
[25] 贺元香,史宝明,张永. 基于本体的语义相似度算法研究[J]. 计算机应用与软件,2013,30(11):312 – 315.
[26] 邓兰兰,李春旺. Web 数据关联创建策略研究[J]. 现代图书情报技术,2011(5):1 – 6.
[27] 敦煌壁画叙词表关联数据查询[EB/OL]. [2021 – 01 – 06]. http://dh.whu.edu.cn/dhvocab/sparql.

作者贡献说明:

高劲松:论文指导、撰写与修改;
付家炜:数据采集、实验设计与操作、论文撰写与修改;
李珂:实验设计与操作、论文撰写。

A Method of Entity Semantic Similarity Calculation for Dunhuang Mural Thesaurus Linked Data with Experiment

Gao Jinsong[1] Fu Jiawei[1] Li Ke[2]
[1] School of Information Management, Central China Normal University, Wuhan 430079
[2] Qingdao Hisense Hitachi Air Conditioning Marketing Co., Ltd, Qingdao 266510

Abstract: [**Purpose/significance**] With the developing of cultural heritage digitization and humanities computing paradigm, the demand of cultural heritage data resources from scholars in the field of humanities have increasingly highlighted when participating in digital humanities research. The semantic integration and interoperability of multi-source and heterogeneous cultural heritage information resources has become a key issue in the construction of digital humanities data infrastructure nowadays, and the effective method of entity semantic similarity calculation has became an important means to achieve this goal. [**Method/process**] Based on the analysis of the ontology model and data framework of Dunhuang Mural Thesaurus Linked Data, this paper proposed an entity semantic similarity calculation method based on the integration of multi granularity matching and weighted calculate, and selected "Feitian" related entities in the dataset as the experimental object to compare the effects of the method proposed in this paper with current methods base on attribute characteristic or edit distance in semantic similarity calculation. [**Result/conclusion**] The experimental results show that, compareing with the other methods, the entity semantic similarity calculation method based on multi-granularity matching can better adapt to the content and structural characteristics of Dunhuang Mural Thesaurus Linked Data, and has better performance in the accuracy of calculation. Thus this paper has introduced another feasible idea for promoting the data interconnection and knowledge sharing of heterogeneous human information resources under the background of digital humanities.

Keywords: Dunhuang murals linked data multi-granularity semantic similarity entity similarity